Audio Engineering Society

Convention Paper

Presented at the 113th Convention
2002 October 5–8 Los Angeles, CA, USA

Perceptually-Based Joint-Program Audio Coding

Christof Faller[1], Raziel Haimi-Cohen[2], Peter Kroon[1], and Joseph Rothweiler[1]

[1] *Media Signal Processing Research, Agere Systems, Murray Hill, New Jersey 07974, USA*

[2] *Advanced Technology Group, Lucent Technologies, Murray Hill, New Jersey 07974, USA*

Correspondence should be addressed to Christof Faller (cfaller@agere.com)

ABSTRACT

Some digital audio broadcasting systems, such as Satellite Digital Audio Radio Services (SDARS), transmit many audio programs over the same transmission channel. Instead of splitting up the channel into fixed bitrate subchannels, each carrying one audio program, one can dynamically distribute the channel capacity among the audio programs. We describe an algorithm which implements this concept taking into account statistics of the bitrate variation of audio coders and perception. The result is a dynamic distribution of the channel capacity among the coders depending on the perceptual entropy of the individual programs. This solution provides improved audio quality compared with fixed bitrate subchannels for the same total transmission capacity. The proposed scheme is non-iterative and has a low computational complexity.

1 INTRODUCTION

Audio signals are usually non-stationary and have a time-varying perceptual entropy [1]. When a perceptual audio coder, such as MPEG-2 AAC [2] or PAC [3], encodes an audio signal at a transparent quality, i.e. with the quantization noise shaped in time and frequency such that it is just below the masked threshold [4], then the bitrate varies from frame to frame. Most broadcasting systems use a constant bitrate transmission channel. To enable variable bitrate transmission over such a channel, the bitstream is buffered and read from the buffer at a constant rate. The buffer must be large enough to absorb the variations in the bitrate. To avoid buffer overflow, a *buffer control* scheme is needed to control the audio coder. Buffer underflow is not a problem because it can be prevented by simply injecting additional bits into the bitstream.

Digital satellite radio [5] such as is being introduced in the United States by Sirius and XM is broadcasting a large number of *radio programs* (about 100) through the same transmission channel.

Previous approaches for joint audio coding [6] heuristically determine the bitrate for each frame of each encoder as a function of the buffer level and the bitrate of the audio coders at masked threshold. Each encoder then uses an iterative algorithm (*rate-loop* [7]) to encode the frame at the specified bitrate. The main disadvantages of these algorithms are that the perceived distortion is not explicitly controlled and their high computational complexity due to the need of a rate-loop.

The scheme we propose in this paper controls the perceived distortion of each encoder explicitly by minimizing the fluctuations of the used *perceptual distortion criterion*. In our proposal we use the *noise-to-masked ratio (NMR)* [8] which is commonly used in most perceptual coders. An important additional advantage of the proposed scheme is that it does not require a rate-loop, which makes its complexity lower than previous approaches.

In Section 2 we present the topology of the joint coding scheme. Also, it is shown why joint coding is better than encoding and transmitting radio programs separately. Section 3 presents a buffer control scheme for joint coding based on statistics and considering perceived distortion [9]. Results are presented in Section 4. Section 5 draws conclusions about the presented scheme and results.

2 TOPOLOGY OF THE JOINT CODING SCHEME

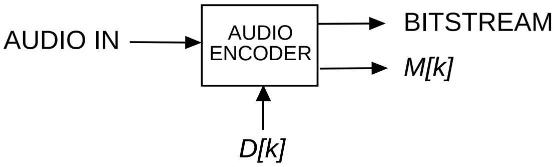

Fig. 1: The *core* of an audio encoder contains a variable bitrate encoder without a bitrate or buffer constraint.

An audio encoder *core*, as shown in Fig. 1, contains an audio encoder without bitrate control or buffer control, e.g. a perceptual audio encoder without a rate-loop. This core encodes each frame k with quantization noise as determined by the perceptual model and a perceptual distortion criterion $D[k]$. The distortion criterion $D[k]$ determines how much quantization noise is introduced above the masked threshold. For $D[k] = 0$ frame k is encoded with quantization noise just below the masked threshold. For $D[k] > 0$ the quantization noise exceeds the masked threshold and as a result fewer bits are needed to encode the frame. However, the resulting audio quality will degrade for increasing values of $D[k]$. Changing the value of $D[k]$ is the most common way to control the resulting bit rate. This mechanism is also used for buffer control purposes where the rate-loop determines $D[k]$ in an iterative manner for each frame.

Joint encoder with identical $D[k]$ values
Figure 2 shows a joint encoder consisting of N identical audio encoders each using the same setting for the distortion value $D[k]$. Having a common distortion criterion is simpler than dealing with a separate distortion criterion for each encoder. With one distortion criterion each encoder is treated equally and all audio programs are encoded at the same quality. The bitrate of each joint frame, $M[k]$, is the sum of the bitrates of the frames of the individual encoders $M_n[k]$ $(n \in \{1, 2, \ldots, N\})$,

$$M[k] = \sum_{n=1}^{N} M_n[k] \,. \qquad (1)$$

Joint Encoder with different $D[k]$ values
Sometimes it is desirable to transmit different audio programs with different quality settings. For example, a satellite radio operator may want to provide a

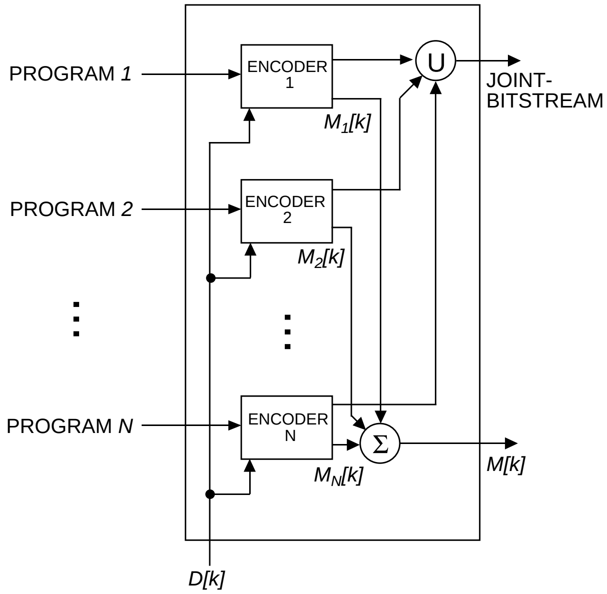

Fig. 2: A joint encoder with identical audio encoders for a similar coding quality for all audio programs. The bitstreams of the N encoders are combined. The bitrate of a joint frame is $M[k]$. A single common distortion criterion is used.

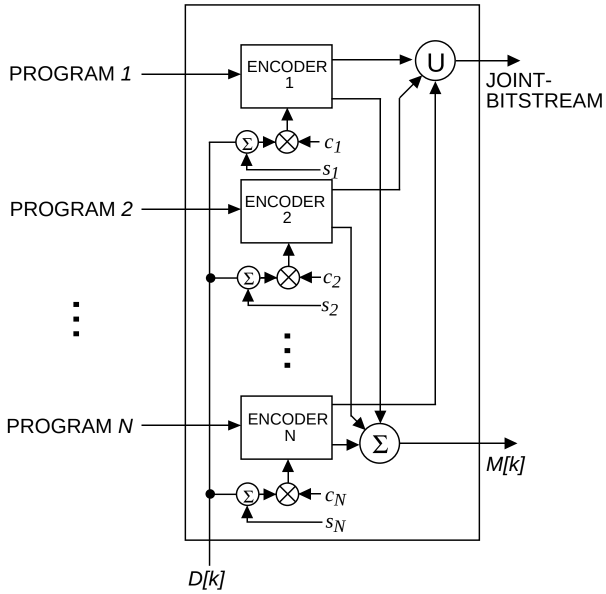

Fig. 3: A joint encoder with different distortion values for the different audio programs. Possibly different audio encoders can be combined. The constants c_i and s_i ($1 \leq i \leq N$) are used to equalize the perceptual impact of $D[k] > 0$.

number of high audio bandwidth stereo music programs and a number of mono news/talk programs with a lower audio bandwidth. In such a scenario an audio encoding algorithm is used with different settings (distortion, audio bandwidth) for the different programs. Even different encoding algorithms may be used for the different kinds of programs. For example, one could use a transform coder for the high quality music channels and a parametric audio coder for the news/talk channels.

A joint encoder with different audio coders operating at different qualities is shown in Fig. 3. The perceptual distortion input of each encoder i is

$$(D[k] + s_i)c_i\,, \qquad (2)$$

where the offset s_i determines the target amount of distortion in relation to the other encoders. The factor c_i is determined such that the relative perceptual impact of $D[k]$ on each encoder is the same. These constants are determined in an empirical way by setting $D[k] = 0$ and adjusting for each encoder s_i the distortion such that the encoder provides the desired audio quality. Then c_i is determined such that all encoders degrade perceptually equally as $D[k]$ increases at a global level.

Joint Encoder Buffer Control

Except for its multiple audio inputs the joint encoders of Figs. 2 and 3 are similar to a single audio encoder core (Fig. 1). Therefore the buffer control of the joint encoder can be designed independently of the number of audio encoders. In fact, the scheme is very similar to the case of only one encoder. A buffered joint encoding scheme with a receiver is shown in Fig. 4. The joint frames of the joint encoder are put into the FIFO *joint buffer.* A *buffer control* scheme determines $D[k]$ such that the buffer level does not overflow. Buffer underflow is not a problem because it can be easily prevented by padding additional (non-used) bits to the frame when underflow would occur. The bits in the joint buffer are transmitted to the receiver with a constant bitrate NR_d. Once a joint frame arrives at the receiver the bits of the radio program P, which is being played back, are extracted and placed into

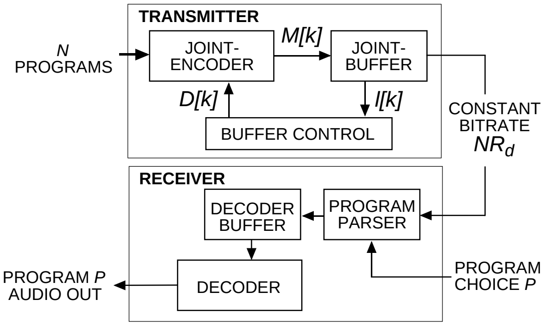

Fig. 4: The scheme of the joint encoder with buffer control is similar to a single coder with buffer control. The bitstream parser at the receiver extracts the bitstream of a specific radio program P.

the *decoder buffer* by the *program parser.*

Compared to a single audio coder the statistics of the momentary joint bitrates $M[k]$ are much more favorable. If we assume that the momentary bitrates of the single audio coders $M_n[k]$ ($n \in \{1, 2, \ldots, N\}$) are random variables with means $m_i = m$ and standard deviations $\sigma_n = \sigma$ then the standard deviation of the joint bitrate $M[k]$ (1) is $\sqrt{N}\sigma$. If the average bitrate available for one audio coder is R_d, then the average bitrate available for the N audio coders is NR_d. The standard deviation of the bitrate normalized to the desired bitrate R_d for a single audio coder is σ/R_d, whereas the standard deviation of the joint encoder, normalized with the available total bitrate NR_d, is reduced to $\sigma/(\sqrt{N}R_d)$. Moreover, in case the mean bitrates m_n of the audio coders are not the same, there is still a significant reduction in the normalized standard deviation for the joint encoder. As a result the joint buffer can be either much smaller than N times the buffer-size of a single audio coder for the same performance, or much better performance can be achieved with the same corresponding buffer sizes.

Another important advantage of joint coding is that the different audio coders can operate at different average bitrates according to how demanding their audio inputs are. This reduces significantly the variations in quality as a function of the input audio material. This is illustrated in Section 4 by comparing joint coding to encoding and transmitting radio programs separately.

3 STATISTICAL BUFFER CONTROL

In the scheme shown in Fig. 4, at each time k, the $M[k]$ bits of the encoded joint frame are put into a FIFO (first-in-first-out) joint buffer while NR_d bits are removed from the joint buffer by the constant bitrate transmission channel. The number of data bits in the joint buffer can be expressed iteratively as

$$l[k] = l[k-1] + M[k] - NR_d, \qquad (3)$$

with the initial buffer level equal to the target buffer level, $l[0] = l_d\ bits$. The task of the buffer control scheme is to monitor the buffer level $l[k]$ and influence the encoding process to make sure the joint buffer does not overflow.

The buffer control scheme we are using is based on the statistical approach presented in [9]. The aim is to vary $D[k]$ as little as possible over time. If no additional delay is allowed for the joint encoding, only frames $k, k-1, k-2, \ldots$ are available at time k. At each time k the average bitrate estimated with $D[k]$ constant within the estimation window $w[i]$ is

$$f_k(d) = \sum_{i=k-W+1}^{k} w[i] M[i]|_{D[k]=d}\,, \qquad (4)$$

where $w[i]$ has a time span of W joint frames. We use a simple rectangular estimation window. The bitrate is estimated for a constant distortion $D[k] = d$ because the aim is to operate the joint encoder while varying $D[k]$ as little as possible. The resulting bitrate estimations are valid for the scenario we are aiming at (constant distortion). Figure 5 shows schematically the function $f_k(d)$. If each joint frame is encoded with a distortion such that the estimated average bitrate is equal to the desired bitrate NR_d,

$$D[k] = f_k^{-1}(NR_d)\,, \qquad (5)$$

then the expected long-term average bitrate of the joint encoder is equal to the desired bitrate.

Next we show why encoding the joint frames according to (5) does not prevent buffer overflow. If we assume that the difference between the frame bitrate and the desired average bitrate,

$$e_M[k] = M[k]|_{D_{NR_d}[k]} - NR_d\,, \qquad (6)$$

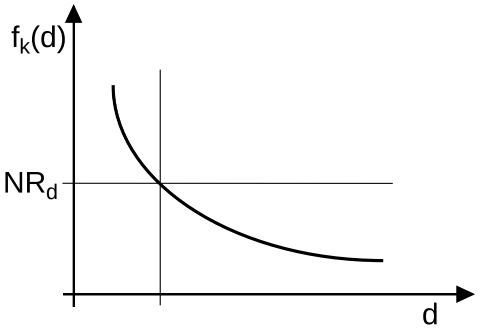

Fig. 5: The function $f_k(d)$ and the point at which the estimated average bitrate is equal to the desired bitrate NR_d.

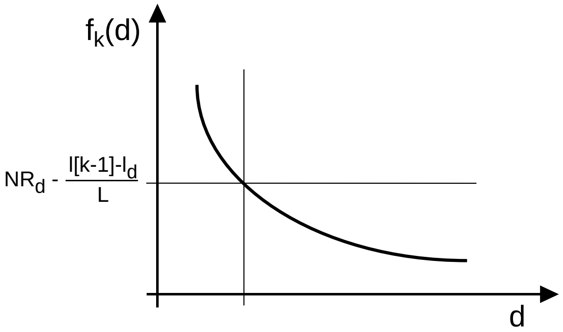

Fig. 6: The function $f_k(d)$ and the point at which the joint frame is encoded.

is an independent and identically distributed (i.i.d.) random variable with a variance of σ_M^2, then the buffer level (3) is the sum of k i.i.d. random variables with a total variance of $k\sigma_M^2$. Because the variance is increasing linearly as a function of time k the buffer will overflow, no matter how large it is, as k goes to infinity.

To operate the joint encoder such that the variance of the buffer level has an upper bound the distortion for each joint frame $D[k]$ is chosen such that the estimated average bitrate $f_k(d)$ (4) is equal to

$$R_{BC}[k] = NR_d - \frac{l[k-1] - l_d}{L}, \tag{7}$$

where L determines how much the coding process is influenced by the buffer level. The subscript $_{BC}$ stands for *buffer control* denoting that with the bitrate $R_{BC}[k]$ buffer control is taken into account.

The corresponding distortion is

$$D[k] = f_k^{-1}(R_{BC}[k])\,. \tag{8}$$

Each joint frame has an expected bitrate of $R_{BC}[k]$ instead of the desired bitrate NR_d. Thus, the buffer level is statistically driven to the desired buffer level l_d with a time constant of LT seconds, where T is the duration of one frame in seconds. For our experiments with a joint encoder with 20 audio coders we chose $L = 250$. With such a large L the distortion $D[k]$ is virtually identical for (5) and (8) except that (8) prevents the buffer level to drift away from its desired level l_d. Figure 6 shows schematically the function $f_k(d)$ (4) and the point at which a joint frame is encoded. Figure 6 is the same as Fig. 5 except that the bitrate on the y-axis is chosen slightly differently.

Next we show that when the joint encoder is operated with distortions according to (8), the mean of the buffer level $E\{l[k]\}$ is l_d and its variance $E\{(l[k] - l_d)^2\}$ is upper bounded by

$$\sigma_e^2 \frac{1}{1 - (1 - \frac{1}{L})^2}\,, \tag{9}$$

where σ_e^2 is $E\{e^2[k]\}$ with

$$e[k] = M[k] - (NR_d - \frac{l[k-1] - l_d}{L}). \tag{10}$$

The variable $e[k]$ is the difference between the frame bitrate and $R_{BC}[k]$ (7) and is assumed to be i.i.d. with zero mean. For the derivation of the mean $E\{l[k]\}$ and the bound for the variance (9), let us re-write the buffer level (3) with (10) as

$$l[k] = 1 - \frac{1}{L} l[k-1] + \frac{1}{L} l_d + e[k]\,. \tag{11}$$

With an initial buffer level of $l[0] = l_d$ and the first joint frame to be encoded $k = 1$, (11) can be written non-iteratively is

$$l[k] = l_d + \sum_{i=1}^{k} e[i](1 - \frac{1}{L})^{k-i}\,. \tag{12}$$

Considering that $e[k]$ has zero mean yields

$$E\{l[k]\} = l_d\,, \tag{13}$$

and the variance as a function of k is

$$E\{(l[k]-l_d)^2\} = \sum_{i=1}^{k} \sigma_e^2 (1-\frac{1}{L})^{2(k-i)} . \qquad (14)$$

Given (14) one can easily show that the variance of the buffer level converges to the value given in (9).

Efficient Implementation

In this section, we describe a scheme for efficient implementation of the proposed joint encoder. For each joint frame k, the buffer control scheme needs to determine $f^{-1}(R_{BC}[k])$(7). This is accomplished by approximating the function $f_k(d)$ via linear interpolation between a set of discrete points. The discrete points are obtained by computing the estimated bitrates $\{R_i[k] = f_k(D_i)\}$ given a set of predefined distortions $\{D_i\}$ (4),

$$R_i[k] = \sum_{i=k-W+1}^{k} w[i]M[i]|_{D_i} , \qquad (15)$$

with $i \in \{1, 2, \ldots, I\}$. Figure 7 shows an example of the approximation of $f_k(d)$ given the discrete points (R_i, D_i). Joint frame k is encoded with a distortion as given by (8) using the estimation of $f_k(d)$.

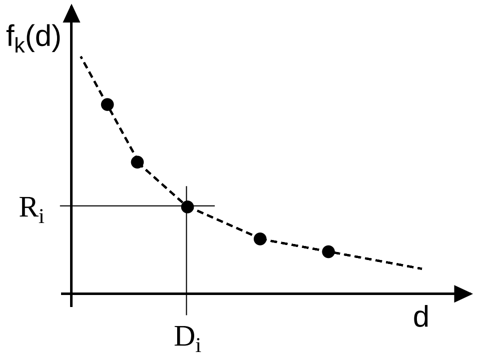

Fig. 7: The function $f_k(d)$ is approximated by interpolating linearly between a few discrete points.

Each audio encoder encodes each frame I times in order to estimate $f_k(d)$. Additionally, each frame is then encoded with a distortion as given in (8). Therefore, the number of coding iterations each audio coder needs to carry out for encoding one frame is

$$I + 1 . \qquad (16)$$

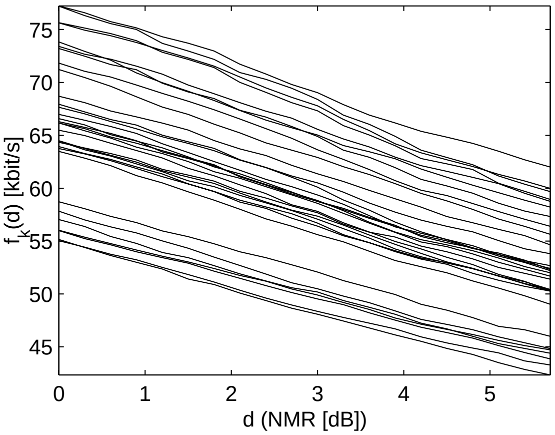

Fig. 8: The function $f_k(d)$ for different frames in different audio clips for PAC.

For PAC, the function $f_k(d)$ plotted for different frames k within different audio signals is shown in Fig. 8. It can be seen that the function $f_k(d)$ can be accurately approximated with a straight line with a time-invariant slope q within the range of operation used. For a joint encoder with any number of N PACs this function can be similarly modeled. Therefore, $f_k(d)$ can be approximated by just computing one point $(D_1, R_1[k])$,

$$f_k(d) \approx q(d - D_1) + R_1[k] . \qquad (17)$$

As a result the PAC encoder needs only to carry out 2 iterations per frame (Eq. 16 with $I = 1$). The presented scheme is significantly less complex than a joint encoder based on traditional rate-loops which usually require more coding iterations.

4 EXPERIMENTAL RESULTS

We compared the performance of a single PAC audio coder with a statistical buffer control scheme as described in [9] to a joint encoder with 20 PAC audio coders operating with the same audio quality (Fig. 2). We compiled 20 different audio programs by concatenating about 30 music and speech clips in random order for each program. The total length of each of the resulting audio programs was about 13 minutes. The 20 audio programs were encoded with the PAC based joint encoder and one audio program was encoded with a stand-alone PAC coder.

The top graph of Fig. 9 shows the distortion $D[k]$ as a function of time for the single and joint encoder. As can be seen in the Fig. 9 the variation in the distortion of the joint encoder is significantly smaller. Additionally, the distortion of the joint encoder stays on average much closer to the masked threshold ($D[k] = 0$). The single audio coder frequently encodes with quantization noise significantly below or above the masked threshold. When the quantization noise is below the masked threshold ($D[k] < 0$) bits are wasted in the sense that more quantization noise could be tolerated. When the quantization noise is above the masked threshold, the encoded audio will be impaired.

The bottom of Fig. 9 shows the bitrate of one specific PAC in the joint encoder. The specific encoder shown encoded the same audio program as the single audio coder. One can see that where the single audio coder is introducing high amounts of quantization noise (top of Fig. 9) the joint coding scheme gives a higher bitrate to the specific coder and thus prevents excessive amounts of quantization noise. This correspondence becomes apparent when comparing the distortion $D[k]$ of the single PAC (top of Fig. 9) with the bitrate of the PAC within the joint encoder (bottom of Fig. 9). These curves are clearly correlated. As a result of this, dependence of the audio quality on the audio material is greatly reduced. For the joint encoder the quality difference of encoded high-demand and low-demand audio clips is much smaller than for a single audio coder.

Figure 10 shows a closeup of the variations in distortion $D[k]$ for the single and joint encoder. The distortion of the joint encoder is much more smooth in time because the normalized standard deviation of the bitrate of the joint frames is much smaller than that of frames of a single audio coder. As a result the perceived quality of the joint encoder will be much more consistent over time.

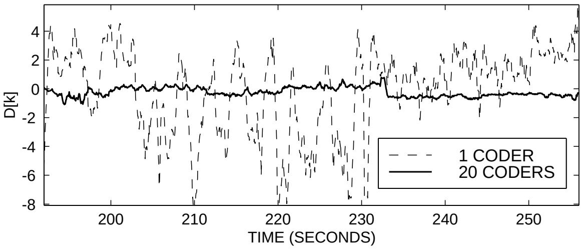

Fig. 10: Short-term variations of the perceptual distortion $D[k]$ for a single audio coder and an audio coder within the joint encoder encoding the same audio material.

5 CONCLUSIONS

A new concept for joint coding of different audio programs has been introduced. The scheme is controlled by a perceptual distortion criterion common to all audio coders within the joint encoder. Therefore, the buffer control affects all coders perceptually equally as opposed to having heuristically occurring stronger degradations on some coders than others. As a result the presented scheme has a more consistent quality among the different radio programs. Also, the variation in the quality over time is reduced. This is different to previous schemes which merely distribute bits among the coders without explicit control over the perceptual impact of the buffer control.

The presented joint coding scheme has a significantly lower complexity than previous schemes which depend on a separate iterative rate-loop for each audio coder. The low complexity made it possible to implement a real-time joint encoding system on low-cost DSP processors.

ACKNOWLEDGMENTS

We would like to thank Frank Baumgarte, Aki Härma, Raymond Jones, Yair Shoham, and Martin Vetterli for the valuable discussions and suggestions.

REFERENCES

[1] J. D. Johnston, “Estimation of perceptual entropy using noise masking criteria,” in *Proc. ICASSP-88*, 1988.

[2] M. Bosi, K. Brandenburg, S. Quackenbush, L. Fielder, K. Akagiri, H. Fuchs, M. Dietz, J. Herre, G. Davidson, and Y. Oikawa, “ISO/IEC MPEG-2 advanced audio coding,” *J. Audio Eng. Soc.*, vol. 45, no. 10, pp. 789–814, 1997.

[3] D. Sinha, J. D. Johnston, S. Dorward, and S. Quackenbush, “The perceptual audio coder

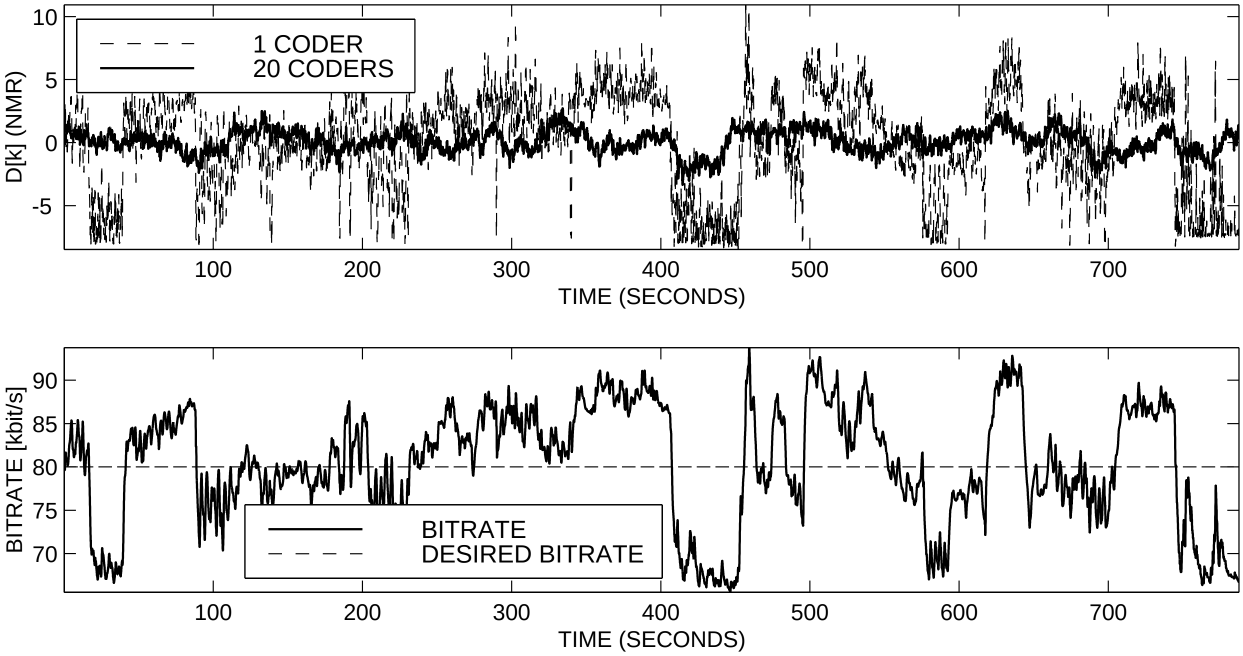

Fig. 9: Top: the perceptual distortion $D[k]$ as a function of time for a single audio coder and an audio coder within a joint encoder with 20 audio coders. Bottom: the bitrate of one audio coder within the joint encoder encoding the same audio material as the single audio coder shown on top.

(PAC),” in *The Digital Signal Processing Handbook*, V. Madisetti and D. B. Williams, Eds., chapter 42. CRC Press, IEEE Press, Boca Raton, Florida, 1997.

[4] M. R. Schroeder, B. S. Atal, and J. L. Hall, “Optimizing digital speech coders by exploiting masking properties of the human ear,” *J. Acoust. Soc. Am.*, vol. 66, no. 6, pp. 1647–1652, Dec. 1979.

[5] R. D. Briskman, “Satellite DAB,” *Int. Journal of Satellite Communications*, vol. 13, pp. 259–266, 1995.

[6] D. Sinha and C.-E. W. Sundberg, “Methods for efficient multiple program digital audio broadcasting,” in *Preprint AES 108th Convention*, Feb. 2000.

[7] F.-R. Jean, C.-F. Long, T.-H. Hwang, and H.-C. Wang, “Two-stage bit allocation algorithm for stereo audio coder,” in *IEEE Proc.-Vis. Image Signal Process.*, Oct. 1996, vol. 143, pp. 331–336.

[8] K. Brandenburg and T. Sporer, “”NMR” and ”Masking Flag”: Evaluation of quality using perceptual criteria,” *Proc. 11th int. AES Conf. Portland, Oregon 1992*, 1992.

[9] C. Faller, “Audio coding using perceptually controlled bitstream buffering,” in *Preprint 111th Conv. Aud. Eng. Soc.*, Sept 2001.